Wei Hu. A Master's Paper for the M.S. in I.S degree. April, 2016. 40 pages. Advisor: Arcot Rajasekar

This project is a part of the DataBridge project, where we try to find similar datasets among a large number of medical datasets stored in the DataBridge server using key words extraction and similarity algorithms. In this project, a sample of 1,000 datasets were randomly chosen from the 18,000 datasets corpus. Modified TF-IDF was used in the sample data to generate key words for the 1,000 datasets and similarity analysis was followed. According to the results, we find that the key words extraction works fine in calculating similarities between different datasets.

Headings:

Information analytics

Term frequency - inverse document frequency

MEDICAL DATA SIGNATURE EXTRACTION USING MODIFIED TF-IDF IN DATABRIDGE PROJECT

by
Wei Hu

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April 2016

Approved by

_____

Arcot Rajasekar

Table of Contents

Introduction 2
Literature review 3
Big data and information analytics 3
Information analytics in health information 4
Improving health care 4
Disease control and prediction 6
Literature review summary 7
Research design 9
Data analysis 10
Data description 10
Keywords (signature) extraction 10
Individual signature approach 11
Corpus signature approach 17
Conclusions and discussion 24
Bibliography 27
Appendix 28
Sample dataset 28
Code for generating term corpus 33
Code for individual signature extraction 35
Code for corpus signature extraction 37

Introduction

Information analytics has been applied to many fields in recent years, and health information analytics is one of the examples. Research has been done in this fields, helping health organizations and hospitals to reduce costs and save people's lives. However, in such field where data sharing and reuse is rather important for the further development in health care in the long run, and due to the complexity of health information data, more research need to be done to find better ways to address the communication and data sharing issues concerning health data.

This study aims to apply information analytics methods to health data and enhance communication and data sharing in health care by identifying similar health datasets for potential health care researches. This study is also a part of DataBridge project, where the specific research question "how can we find more relevant datasets when the data volume is growing rapidly" is expected to be answered. In this study, signature extraction and specific relevant algorithms comparing different datasets were produced to help us determine the similarity between medical datasets. All the datasets were taken from the DataBridge project.

Literature review

Big data and information analytics

Information analytics have been considered to have become increasingly important these days in both the academic and business worlds, within which the ability to understand big data and derive useful information is the key point in this field. Information analytics, is the discovery of useful information in data, and it usually comes with the term “big data”, which is used to describe datasets with large volume and huge complexity (usually with large amount of dimensions compared to traditional datasets). The application of information analytics and big data can be found in many different industries and many studies have highlighted this significant development (Chen, 2012). Perhaps the most well-known areas are e-commerce and market intelligence, where the emergence of customer-generated Web 2.0 technology is used by many companies to improve recommendation system and product assessment for better marketing plans or marketing strategies. With the combination of social media data, customer opinions and satisfaction can be derived by the adoption of text analysis and sentiment analysis techniques (Chen, 2012). The term “e-government and politics” is also an example of adopting similar technology to support online political participation, policy discussions and campaign advertising and other political activities (Chen, 2009)

Information analytics in health information

Health information is also a big part of information analytics, with the fact that the healthcare industry historically has generated very large amounts of data, driven by record keeping, compliance and regulatory requirements, and patient care. And with the development of medical science and health care, we have reasons to believe that the data volume should grow even faster with a dramatically speed. (Raghupathi, 2014). In recent decades, electronic health records (EHR) have been widely adopted in hospitals and clinics not only in the US but also worldwide, which makes it possible to apply information analytics technology. This allows information analytics to make contributions to many more areas such as clinical decision making and patient-centered therapy in hospitals. While it is also worth noting that this will provide benefits outside hospitals. Even genome and environmental issues can enjoy the potential benefits from information analytics (Chen, 2012). One of the recent health big data analytics programs is the National Science Foundation (NSF) Smart Health and Wellbeing (SHB) program, which seeks to address fundamental technical and scientific issues in order to support the development on wellbeing.

Improving health care

While we see the potential of applying information analytics to health care, we also need to face some challenges. Due to its unstructured data with a variety of different terminologies and formats, information analytics in health care has its own complexity and challenges. Also, the electronic health records, as mentioned above, may also bring some privacy issues. Although privacy issue is not a problem existing only in health care,

we do need to pay much attention in this field. However, this research focus more on the complexity challenge.

One strategy to address the complexity challenges is to establish health information organizations (HIOs). Researches have shown that HIOs introduce new ways to improve the efficiency of public health reporting, with even higher quality. This also helps us to gain some insights on how the clinical community should communicate well enough to be able to response properly to emergency issues. And the collaboration also enables the community to conduct health investigation with high quality (Shapiro, 2008). Another approach is to build a platform for analytics using electronic health record data which is called Analytic Information Warehouse (AIW). AIW is able to solve the complexity issue by combining different physical schemas in to a common data structure with derived variables specified to enable the reuse of the data. Another advantage is that AIW derives variables with acceptable correctness, which is very crucial when combining different data schemes. As the literature paper noted, AIW is also able to export the combined data with derived variables into standard forms, which allows many analysis tools to work on the data (Post, 2013).

Many researches and reports have also indicated that using information analytics technology in health care can introduce many benefits to this industry. Some outputs from reports have shown that information analytics has the ability to improve health care processing and optimizing decision-making process. This in turn brings benefits to a large number of hospitals by reducing costs. And the most exciting thing is that many people's lives are saved thanks to the hospital improvements (Raghupathi, 2014). One report from The University of Michigan Health System has also indicated that using

analytics technology can reduce expenses due to the lower rate of transfusions. (Cottle, 2013). Another example is that North York General Hospital with real-time analytics gains greater insight into the operations of healthcare delivery and improved patient outcomes (Cottle, 2013).

There are many other different use cases in which information analytics technology can play an important role, among which managing high-risk patient and predicting readmissions have also attracted attentions from researchers (Bates, 2014).

Disease control and prediction

One of the biggest issue in health information analytics is the disease control. And in many of such cases, social media data plays an important role in this field. Researchers at the Johns Hopkins School of Medicine discovered that they could use data from Google Flu Trends to predict sudden increases in flu-related emergency room visits at least a week before warnings from the CDC. Similarly, the analysis of Twitter updates was as accurate as (and two weeks ahead of) official reports at tracking the spread of cholera in Haiti after the January 2010 earthquake (Cottle, 2013).

In terms of disease perdition in health care. A report from IBM presented a case in which they could predict the likely outcomes of diabetes patients using patients' panel data linked to physicians, management protocols, and the overall relationship to population health management averages (Cottle, 2013). In another example related to diabetes application, physicians at Harvard Medical School and Harvard Pilgrim Health Care recently demonstrated the potential of analytics applications to electronic health records (EHR) data to identify and group patients with diabetes for public health

surveillance. The analytics application also differentiated between Type 1 and Type II diabetes (Cottle, 2013).

Literature review summary

Previous research has indicated the fact that big data analytics has the potential to change the way healthcare providers works and help them to provide better service when applying sophisticated technologies to gain insights from the clinical data repositories. It can also help us with disease prediction and disease control. However, most examples of disease control and process optimization are successful within certain hospitals or organizations. Due to the challenges and complexity mentioned, a research gap does exist on how we can collaborate on sharing health data to enable the reuse of data in a larger scope. HIOs and AIW have given some possible solutions to address this issue, but in the field of disease control and prediction, or some other specific situations where the Ebola disease is spreading and taking people's lives, gathering data from different hospitals and medical organizations turns out to be rather important.

Therefore, more researches need to be done to explore other efficient ways to data collaboration. One possible way to solve this problem is to build a reliable platform where data discovery is supported. In such platform, researchers are able to find relevant datasets for their projects, and the relevance datasets in turn make it possible to do further researches which will benefit the health care field. This requires that we can find good ways to compare different datasets and determine the similarity between them. And when it comes to comparing similarities, we would also want good methods to extract key words from the datasets since it is difficult to compare the whole datasets. Also, the whole datasets may have useless information when doing the similarity analysis. Previous

researches have shown that TF-IDF is a good way to extract keywords from documents. A good example is to extract key information from posts from internet forums and gather similar posts together to have a better support function (Alodadi, 2015). Previous research has also shown that TF-IDF is a good way to extract keywords from micro-blogs (Huang, 2013). Medical datasets can be also viewed as one kind of document since it contains much information in a text format. Therefore, we believe using TF-IDF would also be a possible way to extract keywords in our project when preparing for the similarity analysis.

Research design

This is an information analytics project using the DataBridge platform (Rajasekar, 2013). DataBridge is an indexing mechanism for scientific dataset, and it is similar to current web search engines, DataBridge uses the sociometric analysis to find similar dataset in many reseach fields. DataBridge mainly has three stages, and they are signature generation stage, relevance algorithm stage and sociometric network analysis stage. This study focus on the first stage, where signatures are extracted from the sample datasets.

After the ideal datasets are acquired and signatures are extracted, DataBridge will be applied to find the similarity between different datasets. For example, different datasets representing Ebola cases in different areas or countries in West Africa (mainly in Guinea, Liberia and Nigeria) are expected to have higher similarity scores compared to other datasets representing other diseases. This study would compare the similarity results formed by new relevance metrics with similarity results formed by existing relevance metrics to assess the performance. As mentioned above, the similarity score between some datasets are expected to be larger than other similarity scores, and this would be the basis of relevance metric performance.

The findings of this study are expected to provide insights to the DataBridge project as well as for research in health information, answering “how can we find more relevant health and medical datasets”, which will benefit future researchers to find relevant datasets to support corresponding health information projects.

Data analysis

Data description

All the medical datasets in our project are obtained from the DataBridge server. The datasets are extracted from clinicaltrials.gov and stored as JSON files after parsing the words in the DataBridge server. There are over 18-thousand datasets extracted from the clinicaltrials.gov database. Each dataset contains at least over 40 attributes such as Has Data Monitoring Committee, Investigators ICMJE, Eligibility Criteria ICMJE, Biospecimen, Original Primary Outcome Measures ICMJE and much more. Detailed information can be found in the Appendix. Each attribute (actually also known as keys in the format of JSON file) contains a list of words thanks to the data cleaning done by a member of DataBridge project, which split all the strings into lists of words.

Keywords (signature) extraction

As stated before, all the datasets stored as JSON file stored at least 40 attributes. However, when we consider the keywords extraction, we have the idea that those key words should be informative enough to describe this whole datasets. However, those 40 attributes are not equally informative as we go through the whole datasets. Because these datasets are actually describing clinical trials, after the consideration, we finally chose 9 attributes out of all the attributes, which we believe would be more informative compared to other attributes. Those 9 attributes determined to constructed keywords are:

1) Brief title, 2) Short title, 3) Brief summary, 4) Study arm, 5) Intervention, 6) Study type, 7) Condition, 8) Eligibility and 9) Other study id.

The keywords extraction contains two approaches. In this project, we call the list of keywords as signature. In the first approach, we constructed signatures of each individual datasets, which we called individual signature. In the second approach, we constructed the signature of the whole corpus of the sample datasets, which was actually one single signature considered informative enough to describe all the 1,000 sample datasets. We called the second signature as corpus signature.

Before we conducted the two keywords extraction approaches, the formatting issues and general stop words issues were taken care in our process, this can be seen in the program code in the appendix. For the stop words, the stop words list provided by NLTK was applied in the program. Besides, we considered that these datasets are particularly for medical research, therefore, we also added some medical related words as stop words in the stop words list. Those medical stop words were inspired by online searching and the inspection of datasets in the data preparation step. Some example medical stop words were: patient, hospital, drugs, medicine.

Individual signature approach

In this approach, we extracted the 9 attributes stated above. We treated all the words from the chosen attributes in a medical file as a small corpus of one document. We did it in all the 1,000 sample files and therefore we got 1,000 documents. We then checked each word in the corpus and calculated its document frequency (df), meaning how many documents containing this specific word. With this calculation, we could get a matrix of document frequency shown below:

	document 1	document 2	document 3	document 4	…	document 1,000	df
word 1	0	0	1	1	…	0	23
word 2	0	1	0	0	…	1	50
word 3	1	1	0	0	…	0	6
word 4	0	1	0	0	…	1	10
word 5	1	1	0	0	…	0	200
word 6	1	1	1	0	…	1	345
word 7	1	1	0	0	…	0	2
word 8	0	0	1	0	…	0	800
word 9	0	0	1	1	…	1	66
word 10	1	1	0	0	…	0	92
…	…	…	…	…	…	…	…
word N	0	1	1	0	…	0	678

The matrix shown above was calculated in a Python program coded for this particular project and stored in a JSON file as the output. Notice that the values in the matrix above are for demonstration and are not real data values in this project. The code of the Python program can be found in the appendix.

After computing the document frequency for all the words within the sample of 1,000 documents. We then went back and checked each single document and the words within it, meaning that we then inspected each word in every document to see if this word was informative enough for this document. The value to indicate a word/term being informative was called significance. The smaller the value of significance is, the more informative the word was in a document. The significance was computed by the following formula given a term t:

$$\text{Significance}(t) = -P(t)*\text{mIDF}(t) - [1-P(t)]*\text{mIDF}(\text{not } t)$$

$$*P(t) = \text{df}/\text{total number of document};$$

$$*\text{mIDF}(t) = \log(P(t)); \quad *\text{mIDF}(\text{not } t) = \log(1-P(t))$$

In this formula, P(t) is the probability of a given term t. And computing the "df/total number of document" is considered as the probability of the given term in the whole corpus. And mIDF(t), meaning the modified IDF, which is the log of P(t) with the base of 10.

The idea of this formula is adapted from the idea of the relationship of IDF to entropy. As shown below, given a term t, we can compute its probability and its IDF(t). When we combine the probability and the IDF, we will get a curve graph shown below (Figure 1). This is similar to the typical entropy curve where the two ends of the curve represent the pure collection of the elements, which in our case means the most unique and most general elements while the peak part indicates a chaos mix of different types of elements.

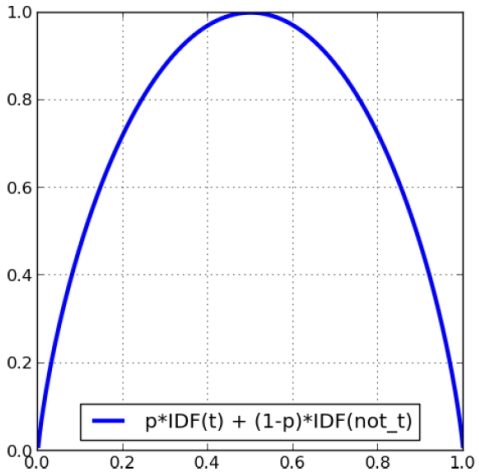

Figure 1. The relationship of IDF to entropy

However, in this project, we used a modified IDF in our formula. The original IDF is equal to the log of the sum of 1 and total number of documents divided by the document frequency. In this project, the modified IDF is equal to the log of probability.

Since the probability in our formula is the reciprocal of the original log, we add a minus sign to the result to make it positive.

The significance using modified IDF ranged from 0 to 0.3. The 0.3 came from the peak part where when P(t) was equal to 0.5, and the both logs of 0.5 were about minus 0.3, multiple by the both P(t) as 0.5. That was how we got the peak value as 0.3. Therefore, we set the threshold as 0.15 and all the words with significance less than 0.15 were considered as informative enough to construct the signature of a single document. In our next step, we went through all the 1,000 files again and extracted signatures for every file. Below is one sample signature taken from the 1,000 computed signatures. The code for computing the signature can be found in the appendix.

["ocular", "-lrb-", "oht", "-rrb-", "reduction", "intraocular", "delay", "onset", "glaucomatous", "visual", "field", "loss", "optic", "disc", "damage", "hypertensive", "judge", "moderate", "develop", "open-angle", "glaucoma", "produce", "natural", "datum", "assist", "identify", "likely", "benefit", "early", "quantify", "among", "individual", "intervention", "Observation", "Close", "commercially", "topical", "hypotensive", "eye", "drop", "man", "nonpregnant", "iop", "hg", "fellow", "best-corrected", "acuity", "worse", "20/40", "life-threatening", "debilitating", "elevated", "angle-closure", "anatomically", "narrow", "angle", "background", "diabetic", "retinopathy", "obscure", "interpretation", "unwillingness", "random", "assignment", "nei-24", "5u10ey009307-16", "5u10ey009341-14"]

With one signature for every single file we had. We then could compute similarity between different files. Particularly, we used cosine similarity for comparing pairs of signature files. First of all, we computed cosine similarities between different files in

their original format. The similarity analysis was done by another project member and some of the results are shown below (Boya, 2016). In the brackets are the two datasets represented by their file names followed by the final value of cosine similarity.

('NCT00000134.json', 'NCT00004562.json', 0.421)

('NCT00000134.json', 'NCT00004563.json', 0.398)

('NCT00000134.json', 'NCT00004635.json', 0.461)

('NCT00000371.json', 'NCT00000378.json', 0.187)

('NCT00000371.json', 'NCT00000392.json', 0.266)

('NCT00000371.json', 'NCT00000479.json', 0.093)

('NCT00000371.json', 'NCT00000575.json', 0.193)

('NCT00000371.json', 'NCT00000620.json', 0.145)

('NCT00000371.json', 'NCT00001151.json', 0.153)

('NCT00000371.json', 'NCT00001213.json', 0.091)

('NCT00000371.json', 'NCT00001566.json', 0.152)

('NCT00000371.json', 'NCT00001586.json', 0.153)

('NCT00000371.json', 'NCT00001596.json', 0.181)

('NCT00000371.json', 'NCT00001656.json', 0.284)

('NCT00000371.json', 'NCT00001703.json', 0.141)

('NCT00000371.json', 'NCT00001723.json', 0.175)

The second step was to apply the signature files to the similarity analysis. Some of the results of pairwise analysis on the signature files are shown below (Boya, 2016). Similar to the results shown above. The pair of signatures are represented by their file names and followed by the value of cosine similarity.

('NCT00003659.json_signature.txt', 'NCT00004500.json_signature.txt', 0.0395)

('NCT00003659.json_signature.txt', 'NCT00004547.json_signature.txt', 0.102)

('NCT00003659.json_signature.txt', 'NCT00004563.json_signature.txt', 0.0958)

('NCT00003782.json_signature.txt', 'NCT00003869.json_signature.txt', 0.0831)

('NCT00003782.json_signature.txt', 'NCT00004092.json_signature.txt', 0.2527)

('NCT00003869.json_signature.txt', 'NCT00004562.json_signature.txt', 0.0796)

('NCT00003869.json_signature.txt', 'NCT00004563.json_signature.txt', 0.0581)

('NCT00003896.json_signature.txt', 'NCT00004092.json_signature.txt', 0.2183)

('NCT00003896.json_signature.txt', 'NCT00004146.json_signature.txt', 0.1)

('NCT00003896.json_signature.txt', 'NCT00004412.json_signature.txt', 0.1048)

('NCT00003896.json_signature.txt', 'NCT00004547.json_signature.txt', 0.201)

('NCT00003910.json_signature.txt', 'NCT00004228.json_signature.txt', 0.1548)

('NCT00004054.json_signature.txt', 'NCT00004635.json_signature.txt', 0.2079)

('NCT00004092.json_signature.txt', 'NCT00004228.json_signature.txt', 0.2143)

('NCT00004092.json_signature.txt', 'NCT00004412.json_signature.txt', 0.1563)

As we can see from the results, the values of cosine similarity of the original files are generally greater than the cosine similarity values of the signature files. However, we also find that some of the values for signature files are very close to the values of the original files. This indicates that the signature files do work well too in some of the files. And we should also note that removing the stop words may be one of the reasons that the similarity values between signature files are generally smaller than the original files, which is acceptable.

Corpus signature approach

After computing the individual signatures for each file. We then considered if we could make use of this sample corpus of 1,000 files and generalize it into the whole 18,000 files. This requires that we generate a more informative signature which contains the key information of this 1,000 files. This signature should be representative enough and can be used as a general signature for other medical or clinical data. After all, this project aims to explore potential similar datasets in a broad scope within the medical research area. Therefore, we also conducted the second approach which was the corpus signature approach. And we wanted the signature contains about 50-100 words.

Based on the document frequency matrix, we then had all the df values of all the terms in the corpus. Since we wanted to generalize this signature to other medical datasets, we would want the words in the signature be general enough to describe what we would expect to see in a clinical dataset. Therefore, we computed the IDF(t) of the given term t to see if the term was appropriate for the corpus signature. Because when computing IDF(t), the smaller the value is, the more general the term is. When we set the threshold to 1.25, a total of 50 words was extracted and shown below.

[‘one', ‘prior', ‘test', ‘year', ‘intervention', ‘month', ‘experimental', ‘-lrb-', ‘Criteria', ‘disease', ‘week', ‘include', ‘1', ‘Comparator', ‘2', ‘drug', ‘-rrb-', ‘trial', ‘criterion', ‘treatment', ‘Inclusion', ‘4', ‘follow', ‘use', ‘6', ‘time', ‘history', ‘within', ‘Exclusion', ‘dose', ‘therapy', ‘may', ‘18', ‘least', ‘purpose', ‘patient', ‘name', ‘age', ‘receive', ‘clinical', ‘3', ‘interventional', ‘treat', ‘active', ‘study', ‘consent', ‘pregnant', ‘day', ‘must', ‘5']

However, in this list we can see that some words are not informational as we expected, such as some numbers and words like “name”, “age” and “study”. After

consideration, we then set the threshold as 1.5, and we got a list of 99 words shown below.

[‘one', ‘prior', ‘and/or', ‘previous', ‘test', ‘year', ‘intervention', ‘month', ‘safety', ‘contraception', ‘experimental', ‘-lrb-', ‘Criteria', ‘disease', ‘female', ‘week', ‘include', ‘1', ‘Comparator', ‘2', ‘surgery', ‘condition', ‘status', ‘drug', ‘-rrb-', ‘evidence', ‘trial', ‘criterion', ‘treatment', ‘evaluate', ‘limit', ‘give', ‘medical', ‘Inclusion', ‘skin', ‘cancer', ‘4', ‘investigational', ‘blood', ‘significant', ‘mg', ‘require', ‘medication', ‘follow', ‘woman', ‘use', ‘infection', ‘chemotherapy', ‘pregnancy', ‘s’, ‘effective', ‘6', ‘renal', ‘time', ‘history', ‘within', ‘Exclusion', ‘dose', ‘diagnosis', ‘therapy', ‘subject', ‘may', ‘10', ‘12', ‘18', ‘two', ‘least', ‘purpose', ‘patient', ‘30', ‘greater', ‘creatinine', ‘name', ‘disorder', ‘age', ‘receive', ‘less', ‘phase', ‘clinical', ‘3', ‘cell', ‘control', ‘interventional', ‘treat', ‘effect', ‘active', ‘study', ‘potential', ‘daily', ‘normal', ‘consent', ‘pregnant', ‘day', ‘chronic', ‘must', ‘determine', ‘5', ‘Drug', ‘know']

The idea to change the threshold is that, we would like to get a bigger list of keywords at first. Then we can see all the potential words in the list and determine if every word is informative enough. If the word is not informative, we can delete it from the list. After filtering out some words manually, we can get a more accurate signature for the whole corpus. Meanwhile, the words removed from the list are not informative enough, so we can consider them as stop words. As stated before, we developed our own stop words for medical data without prior knowledge, but now with all the removed words, we can again develop our medical stop words list with more confidence. So getting a bigger list of words would benefit us in two aspects, naming the signature extraction and stop words construction. However, we thought a list of 99 words was not

big enough for our manual filtering, so we changed the threshold again and set it as 1.6. We then got a list of 135 words shown below, which is a much bigger list of words.

[‘one', ‘past', ‘prior', ‘and/or', ‘investigator', ‘previous', ‘participation', ‘test', ‘year', ‘intervention', ‘month', ‘safety', ‘contraception', ‘experimental', ‘hour', ‘primary', ‘-lrb-', ‘Criteria', ‘disease', ‘placebo', ‘female', ‘would', ‘week', ‘without', ‘include', ‘1', ‘Comparator', ‘current', ‘positive', ‘2', ‘take', ‘surgery', ‘condition', ‘status', ‘drug', ‘count', ‘-rrb-', ‘evidence', ‘trial', ‘criterion', ‘treatment', ‘evaluate', ‘exclusion', ‘limit', ‘give', ‘medical', ‘Inclusion', ‘iv', ‘skin', ‘cancer', ‘4', ‘investigational', ‘compare', ‘blood', ‘significant', ‘per', ‘every', ‘mg', ‘require', ‘medication', ‘follow', ‘woman', ‘use', ‘infection', ‘chemotherapy', ‘pregnancy', ‘s’, ‘sign', ‘effective', ‘malignancy', ‘6', ‘total', ‘negative', ‘heart', ‘renal', ‘severe', ‘time', ‘history', ‘platelet', ‘Placebo', ‘within', ‘Exclusion', ‘dose', ‘diagnosis', ‘therapy', ‘subject', ‘failure', ‘may', ‘10', ‘12', ‘18', ‘either', ‘two', ‘least', ‘purpose', ‘patient', ‘30', ‘greater', ‘procedure', ‘creatinine', ‘name', ‘disorder', ‘tumor', ‘complete', ‘age', ‘receive', ‘less', ‘phase', ‘clinical', ‘upper', ‘3', ‘cell', ‘control', ‘interventional', ‘treat', ‘effect', ‘active', ‘study', ‘potential', ‘daily', ‘agent', ‘normal', ‘consent', ‘pregnant', ‘confirm', ‘day', ‘define', ‘chronic', ‘oral', ‘must', ‘determine', ‘5', ‘serum', ‘Drug', ‘know']

We then removed the useless words from this list and finally a list of 59 words was obtained as shown below.

['prior', ‘investigator', ‘intervention', ‘safety', ‘contraception', ‘experimental', ‘criteria', ‘placebo', ‘comparator', ‘current', ‘positive', ‘surgery', ‘condition', ‘status', ‘drug', ‘count', ‘evidence', ‘trial', ‘criterion', ‘treatment', ‘evaluate', ‘limit', ‘skin', ‘cancer', ‘investigational', ‘blood', ‘significant', ‘medication', ‘infection', ‘chemotherapy',

‘pregnancy', ‘effective', ‘malignancy', ‘heart', ‘renal', ‘severe', ‘history', ‘platelet', ‘placebo', ‘dose', ‘diagnosis', ‘therapy', ‘failure', ‘procedure', ‘creatinine', ‘disorder', ‘tumor', ‘clinical', ‘cell', ‘control', ‘interventional', ‘effect', ‘active', ‘potential', ‘pregnant', ‘chronic', ‘oral', ‘determine', ‘serum']

However, when manually filtering out the useless words, or considered as stop words, we found one interesting thing about this process. There were actually two types of stop words in this list. The first type of words were actually stop words with no specific meanings for working as a signature. Those words were usually numbers, and words with less information related to clinical data. A sample of this type of words is shown below.

[past, and/or, per, every, either, name, less, must, use, mg, would, least, give, medical, inclusion, require, exclusion, purpose, complete, receive, phase, consent, day, define, know, exclusion, follow, total, subject, compare, sign, take, participation, daily, confirm, disease, hour, study, treat, primary, normal, patient, previous, greater]

The second type of stop words were actually semi-informative words. On the one hand, they were not informative as actual signature words. However on the other hand, this type of words could be informative in some specific situation. “Female” is one of such semi-informative words. When we consider this word alone, it cannot provide much information in a clinical dataset. However, researches in women health may find this word informative when clinical data about females are in need. Therefore, we thought this type of words were semi-informative words.

After the corpus signature was obtained, we then conducted the similarity analysis using this signature. But we didn’t apply the cosine similarity as we did in the individual

approach. The way we did the analysis was that we compared the signature with each individual signature we had in the first approach. We then computed the number of common words between the corpus signature and the individual signature. If there was 1 common word, we determined the similarity as 0.1. If there were N common words between the pair, we determined the similarity as N/10, where N was less than 10. However, if there were more than 10 common words, we determined the similarity as 1. Below are the analysis results on 48 file pairs done by another project member after the corpus signature approach (Boya, 2016). The corpus signature was named “keywords.txt” and the lowest and highest similarity scores are bolded.

('keywords.txt', 'NCT00000125.json_signature.txt', 0.1)

('keywords.txt', 'NCT00000134.json_signature.txt', 0.3)

('keywords.txt', 'NCT00000371.json_signature.txt', 0.2)

('keywords.txt', 'NCT00000378.json_signature.txt', 0.1)

('keywords.txt', 'NCT00000392.json_signature.txt', 0.2)

('keywords.txt', 'NCT00000479.json_signature.txt', 0.4)

('keywords.txt', 'NCT00000575.json_signature.txt', 0.3)

('keywords.txt', 'NCT00000620.json_signature.txt', 0.2)

('keywords.txt', 'NCT00001151.json_signature.txt', 0.1)

('keywords.txt', 'NCT00001213.json_signature.txt', 0.1)

('keywords.txt', 'NCT00001566.json_signature.txt', 0.5)

('keywords.txt', 'NCT00001586.json_signature.txt', 0.3)

('keywords.txt', 'NCT00001596.json_signature.txt', 0.3)

('keywords.txt', 'NCT00001656.json_signature.txt', 0.1)

('keywords.txt', 'NCT00001703.json_signature.txt', 0.3)

('keywords.txt', 'NCT00001723.json_signature.txt', 0.3)

('keywords.txt', 'NCT00001832.json_signature.txt', 0.6)

('keywords.txt', 'NCT00001941.json_signature.txt', 0.4)

('keywords.txt', 'NCT00001959.json_signature.txt', 0.0)

('keywords.txt', 'NCT00001962.json_signature.txt', 0.3)

('keywords.txt', 'NCT00001984.json_signature.txt', 0.3)

('keywords.txt', 'NCT00002540.json_signature.txt', 0.7)

('keywords.txt', 'NCT00002850.json_signature.txt', 0.2)

('keywords.txt', 'NCT00002975.json_signature.txt', 0.4)

('keywords.txt', 'NCT00003138.json_signature.txt', 0.3)

('keywords.txt', 'NCT00003222.json_signature.txt', 0.6)

('keywords.txt', 'NCT00003224.json_signature.txt', 0.2)

('keywords.txt', 'NCT00003298.json_signature.txt', 0.3)

('keywords.txt', 'NCT00003377.json_signature.txt', 0.2)

('keywords.txt', 'NCT00003389.json_signature.txt', 0.2)

('keywords.txt', 'NCT00003590.json_signature.txt', 0.3)

('keywords.txt', 'NCT00003659.json_signature.txt', 0.3)

('keywords.txt', 'NCT00003782.json_signature.txt', 0.3)

('keywords.txt', 'NCT00003869.json_signature.txt', 0.2)

('keywords.txt', 'NCT00003896.json_signature.txt', 0.4)

('keywords.txt', 'NCT00003907.json_signature.txt', 0.2)

('keywords.txt', 'NCT00003910.json_signature.txt', 0.3)

('keywords.txt', 'NCT00004054.json_signature.txt', 0.4)

('keywords.txt', 'NCT00004092.json_signature.txt', 0.6)

('keywords.txt', 'NCT00004143.json_signature.txt', 0.4)

('keywords.txt', 'NCT00004146.json_signature.txt', 0.1)

('keywords.txt', 'NCT00004228.json_signature.txt', 0.3)

('keywords.txt', 'NCT00004412.json_signature.txt', 0.6)

('keywords.txt', 'NCT00004500.json_signature.txt', 0.1)

('keywords.txt', 'NCT00004547.json_signature.txt', 0.6)

('keywords.txt', 'NCT00004562.json_signature.txt', 0.4)

('keywords.txt', 'NCT00004563.json_signature.txt', 0.1)

('keywords.txt', 'NCT00004635.json_signature.txt', 0.4)

As we can see from the analysis results, we values range from 0.0 to 0.7. While there was only one value of 0.0, nearly every individual signature file had common words with this corpus signature. This indicates that the corpus signature do work well in presenting medical dataset characteristics as an effective signature in a broad scope.

Conclusions and discussion

This project focused on the keywords (signature) extraction from the clinical datasets. Two approaches were tried in the signature extraction: individual signature and corpus signature. In the individual signature, a modified TF-IDF formula was used in the signature extraction. The results of the individual signature extraction showed a good result when applying the similarity algorithms. However, it is also worth noting that we do observe the average value of cosine similarity dropped from the original file comparison to the signature file comparison. Although filtering out the stop words would possibly be the main reason for it, we would like to see future studies on the signature to address a better way to improve the cosine similarity while using the informative words as signatures. Potential study can focus both on the signature obtained in this study to see how the signature interact with each other, or the discovery of better algorithms to extract more informative signatures without losing much similaritiy.

In the corpus signature extraction, we found a more general signature for the whole sample corpus, and the similarity analysis indicated that this signature worked well when applying to other datasets. The way we extracted the signature was using simple IDF values. Even though the results indicates that using IDF was a good approach, we now do wonder if there are other algorithms we can use to improve the performance of signature extraction. In the first approach we used a modified TF-IDF, and we would like to see how we can modify the IDF in the second approach. This will be a good way to

explore in future studies. Besides, the whole process of removing stop words has been a great learning opportunity for future information analytics projects. As we can see from the process, both the NLTK stop words list and a manually developed list were used in this project. This has indicated that in an information analytics project, the combination of algorithms and human intelligence are both important when analyzing the data. All the information analytics projects are aimed to solve a particular real world problem and we believe we need a deep understanding of the project as we making the progress. The algorithms are essential tools for us, but they are not enough as we can see from our project, especially when in an exploratory work like DataBridge. Therefore, this project is a good example concerning how we can use our intelligence in data understanding.

However, there are still two questions need to be answered in the future research concerning the corpus signature we extracted in the second approach. The first question is about the two type of stop words mentioned above. While the first type of stop words should be removed in any data, the second type of stop words, which we call semi-informative words, requires more considerations. Future study can work on this type of semi-informative words to see how we can apply these words to certain research fields in the future, by which we can explore the value of these semi-informative words.

The second question is about the corpus signature. We wonder how this signature would work when applying to datasets outside of DataBridge project. In this study, we only used the data from our own server, so we probably can get some good similarity results. But this project aims to foster more medical studies in the future. One way to assess this signature is to ask medical professionals. As we discussed above, human

intelligence is also important in an information analytics project, so we believe that professional opinions and input would make this project more powerful in the future.

Bibliography

Alodadi, M., & Janeja, V. P. (2015, October). Similarity in Patient Support Forums Using TF-IDF and Cosine Similarity Metrics. In Healthcare Informatics (ICHI), 2015 International Conference on (pp. 521-522). IEEE.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Affairs, 33(7), 1123-1131.

Boya, Harika. (2016). Finding similarity using metadata of clinical trials using Natural Language Processing in DataBridge.

Chen, H. 2009. "AI, E-Government, and Politics 2.0," IEEE Intelligent Systems (24:5), pp. 64-67.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS quarterly, 36(4), 1165-1188.

Cottle, M., Kanwal, S., Kohn, M., Strome, T., & Treister, N. Transforming health care through big data. Strategies for leveraging big data in the health care industry. New York: Institute for Health Technology Transformation; 2013.

Huang, X., & Wu, Q. (2013, October). Micro-blog commercial word extraction based on improved TF-IDF algorithm. In TENCON 2013-2013 IEEE Region 10 Conference (31194) (pp. 1-5). IEEE.

Post, A. R., Kurc, T., Cholleti, S., Gao, J., Lin, X., Bornstein, W., & Saltz, J. H. (2013). The Analytic Information Warehouse (AIW): A platform for analytics using electronic health record data. Journal of biomedical informatics, 46(3), 410-424.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health Information Science and Systems, 2(1), 3.

Rajasekar, Arcot, Hye-Chung Kum, Merce Crosas, Jonathan Crabtree, Sharlini Sankaran, Howard Lander, Thomas Carsey, Gary King, and Justin Zhan. "The DataBridge." SCiENCE 2, no. 1 (2013): pp-1.

Shapiro JS, Mostashari F, Hripcsak G, Soulakis N & Kuperman G. (2008). Using health information exchange to improve public health. J Public Health, 2011(101), 616-623.

Appendix

Sample dataset

"TABULAR_VIEW_MAP_JSON": {
"Has Data Monitoring Committee": "No",
"Investigators ICMJE": "",
"Removed Location Countries": null,
"Eligibility Criteria ICMJE": "Inclusion Criteria: Subjects that sign the Informed Consent form required for prospectively enrolling patients into the study. Subjects that present at a hospital, clinic, or physician\u0027s office with the signs and symptoms of a respiratory tract infection. Subjects with an acute respiratory infection where said acute respiratory infection is suspected of being caused by an Influenza virus. Exclusion Criteria: Subjects where the duration of the symptoms of such an acute respiratory infection is greater than or equal to 5 days (i.e., ≥5).",
"Biospecimen": "Retention: Samples With DNA Description: Extracted nucleic acid, Residual Universal Transport Medium",
"Administrative Information": null,
"Original Primary Outcome Measures ICMJE (submitted: February 18, 2011)": "Detection of Respiratory Viruses [Time Frame: Specimens will be taken within 5 days of the appearance of symptoms.] [Designated as safety issue: No] QIAGEN ResPlex II Advanced Panel are: To establish that the clinical sensitivity and specificity are substantially equivalent to viral culture To establish that the clinical sensitivity and specificity are substantially equivalent to the respective validated nucleic acid amplification-based (i.e., PCR) laboratory developed test (PCR-LDT) artus Influenza A/B RT-PCR Test is: 1.To establish that the clinical sensitivity and specificity are substantially equivalent to standard viral culture",

"Current Primary Outcome Measures ICMJE (submitted: May 22, 2012)": "Detection of Respiratory Viruses [Time Frame: Specimens will be taken within 5 days of the appearance of symptoms.] [Designated as safety issue: No] The presence of Influenza A or Influenza B virus.",

"Collaborators ICMJE": "",

"Study Sponsor ICMJE": "QIAGEN Gaithersburg, Inc",

"Start Date ICMJE": "February 2011",

"Information Provided By": "QIAGEN Gaithersburg, Inc",

"Official Title ICMJE": "Testing of Respiratory Specimens for the Validation of the QIAGEN ResPlex II Advanced Panel Test and the Artus Influenza A/B RT-PCR Test",

"Change History": "Complete list of historical versions of study NCT01302418 on ClinicalTrials.gov Archive Site",

"Brief Summary": "The study will be conducted using nasopharyngeal swab specimens collected prospectively from individuals suspected of having the signs and symptoms of an acute respiratory tract infection caused by a respiratory virus. A series of standard viral culture tests validated for routine use in the clinical laboratory, and/or a series of PCR-based Laboratory Developed Tests (PCR-LDT) validated by a central reference laboratory will be used to verify the performance of the investigational artus Influenza A/B RT-PCR test and the QIAGEN ResPlex II Advanced Panel test. From each specimen five (5) aliquots will be prepared: (a) one aliquot will be tested in real-time using the assigned viral culture reference methods; (b) one aliquot will be used to extract nucleic acid in real-time for investigational testing; (c) one aliquot of the specimen will be stored at --70C for subsequent shipment to the reference laboratory for PCR-LDT testing, (d) one aliquot will be archived at -70C for subsequent follow-up by the reference laboratory (e.g., bi-directional sequencing of positive specimens), and (e) any remaining specimen will be stored for the Fresh vs. Frozen Study. The extracted nucleic acid generated from the second aliquot (i.e., \"b\" above) will be split and subjected to testing by both the artus Influenza A/B RT-PCR test and the ResPlex II Advanced Panel test.",

"Study Design ICMJE": "Observational Model: Case-Only Time Perspective: Prospective",

"Brief Title ICMJE": "Collection and Testing of Respiratory Samples",

"Other Study ID Numbers ICMJE": "C10-INFLUENZA-001",
"NCT Number ICMJE": "NCT01302418",
"Recruitment Information": null,
"Verification Date": "May 2012",
"Ages": "",
"Original Secondary Outcome Measures ICMJE": "Not Provided",
"Study Population": "The study population includes individuals having the signs and symptoms of an acute respiratory tract infection suspected of being caused by a respiratory virus.",
"Completion Date": "November 2011",
"Intervention ICMJE": "Device: artus Influenza A/B RT-PCR Test The investigational assay, used for detecting the presence of Influenza A/B.",
"Primary Completion Date": "July 2011 (final data collection date for primary outcome measure)",
"Listed Location Countries ICMJE": "United States",
"Enrollment ICMJE": "272",
"Tracking Information": null,
"Contacts ICMJE": "Contact information is only displayed when the study is recruiting subjects",
"Recruitment Status ICMJE": "Completed",
"Current Other Outcome Measures ICMJE": "Not Provided",
"Sampling Method": "Non-Probability Sample",
"Study Group/Cohort (s)": "Symptomatic Individuals with signs and symptoms of an acute respiratory tract infection where it is suspected that such signs and symptoms are caused by a respiratory virus infection. Intervention: Device: artus Influenza A/B RT-PCR Test",
"Target Follow-Up Duration": "Not Provided",
"Condition ICMJE": "QIAGEN ResPlex II Advanced Panel Influenza A Influenza B Respiratory Syncytial Virus Infections Infection Due to Human Parainfluenza Virus 1 Parainfluenza Type 2 Parainfluenza Type 3 Parainfluenza Type 4 Human Metapneumovirus A/B Rhinovirus Coxsackie Virus/Echovirus Adenovirus Types B/C/E

Coronavirus Subtypes 229E Coronavirus Subtype NL63 Coronavirus Subtype OC43 Coronavirus Subtype HKU1 Human Bocavirus Artus Influenza A/B RT-PCR Test Influenza A, Influenza B,",

"Detailed Description": "Each year the morbidity and mortality associated with acute respiratory tract infections fluctuates seasonally. This rise and fall is associated with the changing prevalence of respiratory viruses in the population. Myriad respiratory viruses are responsible for these infections. For example, Influenza Virus, Respiratory Syncytial Virus (RSV), Parainfluenza Virus, Human Metapneumovirus, Rhinovirus, and Adenovirus have all been identified as causing such acute infections. Numerous pathogenic subtypes have been identified within most of these viral groups. The outbreak of Severe Acute Respiratory Syndrome (SARS) in 2003 was eventually identified as a Coronavirus; the mortality of SARS among the elderly can be as high as 50%. More recently, Human Bocavirus (HBoV) has also been identified as causing acute respiratory tract infections. In 2005 the HBoV was identified by molecular testing and was found to be the only virus identified in a subpopulation of patients suffering from respiratory tract infections. Apart from supportive measure (e.g., bed rest, hydration, etc.), there are no effective treatments for many of these viral infections; however, antiviral agents (e.g., the neuraminidase inhibitors oseltamivir or zanamivir) can be used to alleviate the severity of flu-like symptoms. Identification of a respiratory virus as the causative agent is important because it eliminates the need for treatment with antibiotics; physicians typically wait 7-10 days for symptoms to alleviate before prescribing antibiotics due to risks associated with exacerbating bacterial antibiotic resistance. Each year the virus population fluctuates, and with it the antigenic presentation of the dominant strains that circulate through the population. Epidemics arise when larger and larger portions of the population do not have innate or acquired immunological resistance to such strain(s) in a given season. The World Health Organization (WHO) maintains a separate website dedicated to tracking outbreaks of influenza, especially avian influenza (https://www.who.int/fluvirus_tracker). These zoonotic transmissions that further adapt to enable human-to-human transmission are of the greatest concern because it is predicted that virtually all humans will be immunologically naïve. Zoonotic transmissions in the human population are monitored in the hope that a pandemic similar to the Spanish Flu of

1918 can be avoided; it is estimated that well over 25 million people died from the Spanish Flu. The United States government also maintains a separate website with resources regarding the flu and pandemic related information (http://www.pandemicflu.gov/). On June 11, 2009 the WHO raised the pandemic threat level to 6 in response to the global appearance of a new strain of swine Influenza A (subtype H1N1). The rapidity with which the H1N1 virus has spread exemplifies the notion that quickly and accurately identifying a viral pathogen associated with an outbreak is critical to global public health. In addition to the threat of an influenza outbreak, the expansion in the number of viruses that cause acute respiratory tract infections compounds the difficulty in correctly and rapidly identifying the primary pathogen; each new virus or subtype increases the complexity of testing. Molecular diagnostic assays are ideally suited to address this complexity. Assays based on the polymerase chain reaction (PCR) can incorporate multiple primers and probes (e.g., multiplexed) in a single reaction to deal with this complexity.4 Such assays are extremely sensitive, have a high degree of specificity, and can be performed very quickly. The artus Influenza A/B RT-PCR test is a real-time PCR assay for the detection and identification of Influenza A and B, while the QIAGEN ResPlex II Advanced Panel test is a nucleic acid amplification-based assay for the detection and identification of a broad range of some of the most common respiratory viruses associated with acute respiratory tract infections. In the present study respiratory specimens will be prospectively collected and tested using the artus Influenza A/B RT-PCR test and the QIAGEN ResPlex II Advanced Panel test.",

"Descriptive Information": null,
"Accepts Healthy Volunteers": "No",
"Publications *": "Not Provided",
"Gender": "Both",
"Current Secondary Outcome Measures ICMJE": "Not Provided",
"Last Updated Date": "May 22, 2012",
"First Received Date ICMJE": "February 18, 2011",
"Responsible Party": "QIAGEN Gaithersburg, Inc",
"Original Other Outcome Measures ICMJE": "Not Provided",

```
  "Study Type  ICMJE": "Observational"
 }
```

Code for generating term corpus

```
import json
import os
import string
from nltk.corpus import stopwords

#load the file
file_path = "H:\\Dropbox\\UNC\\MasterPaper\\Databridge\\Clinicaltrial_data\\newdata"
file_folder = [];
data_folder = [];

#store all the files in file_folder
for filename in os.listdir(file_path):
    open_path = file_path+'\\'+filename;

    with open(open_path) as json_file:
        json_data = json.load(json_file);
        file_folder.append(json_data);

#deal with the \xa0 in keys
for file in file_folder:
    data_folder.append(file["TABULAR_VIEW_MAP_LEMMATIZED_JSON"]);

for data in data_folder:
    keys = data.keys();
    for key in keys:
        new_key = key.replace(u'\xa0',"");
        data[new_key] = data.pop(key);
```

```
# construct medical stopwords
stop = stopwords.words('english');
medical_stopwords = ["patients","medicine","for","four","drugs","hospital"]
stop.extend(medical_stopwords);
stop.extend(string.punctuation);

corpus = dict();
for index, data in enumerate(data_folder):
    print index;
    tf = dict();
    words = data["Brief Title ICMJE"];
    words.extend(data["Brief Summary"]);
    words.extend(data["Study Arm (s)"]);
    words.extend(data["Intervention ICMJE"]);
    words.extend(data["Study Type ICMJE"]);
    words.extend(data["Condition ICMJE"]);
    words.extend(data["Eligibility Criteria ICMJE"]);
    words.extend(data["Other Study ID Numbers ICMJE"]);
    print "done"

    removed_words = [i for i in words if i not in stop];
    for word in removed_words:
        if word not in tf:
            tf[word] = 1;
        else:
            tf[word] += 1;

    for key in tf.keys():
        if key not in corpus:
            corpus[key] = 1;
        else:
```

```
            corpus[key] += 1;

print corpus;
print len(corpus);

#Save the corpus as a json file
with open('corpus.json', 'w') as fp:
    json.dump(corpus, fp);
```

Code for individual signature extraction

```
import json
import math
import os
import string
from nltk.corpus import stopwords

def main():
    file_path =
"H:\\Dropbox\\UNC\\MasterPaper\\Databridge\\Clinicaltrial_data\\newdata";
    corpus_path = "H:\\Dropbox\\UNC\\MasterPaper\\Databridge\\corpus.json";

    stop = stopwords.words('english');
    medical_stopwords = ["patients","medicine","for","four","study"]
    stop.extend(medical_stopwords);
    stop.extend(string.punctuation);

    with open(corpus_path) as corpus_file:
        corpus = json.load(corpus_file);

    for filename in os.listdir(file_path):
        signature = [];
```

```
print filename;
open_path = file_path+'\\'+filename;
with open(open_path) as json_file:
    json_data = json.load(json_file);
    extracted_data = json_data["TABULAR_VIEW_MAP_LEMMATIZED_JSON"];
    keys = extracted_data.keys();
    for key in keys:
        new_key = key.replace(u'\xa0',"");
        extracted_data[new_key] = extracted_data.pop(key);

words = extracted_data["Brief Title ICMJE"];
words.extend(extracted_data["Brief Summary"]);
words.extend(extracted_data["Study Arm (s)"]);
words.extend(extracted_data["Intervention ICMJE"]);
words.extend(extracted_data["Study Type ICMJE"]);
words.extend(extracted_data["Condition ICMJE"]);
words.extend(extracted_data["Eligibility Criteria ICMJE"]);
words.extend(extracted_data["Other Study ID Numbers ICMJE"]);
print "extraction done";

removed_words = [i for i in words if i not in stop];
removed_words = [i for i in removed_words if not i.isdigit()]

sum_of_docs = 1000;

for word in removed_words:
    sig = compute_sig(corpus[word], sum_of_docs);
    if (sig<0.15 and word not in signature):
        signature.append(word);
    else:
        signature;
```

```
        print signature;
        save_name = filename+"_signature.txt";
        with open(save_name, 'w') as fp:
            json.dump(signature, fp);

def compute_sig(df, sum_of_docs):
    sig=1;
    if df==1000:
        sig=1;
    else:
        df = float(df);
        sumdoc = float(sum_of_docs);
        pt = df/sumdoc;
        pnt = 1-pt;
        idft_r = math.log10(df/sumdoc);
        idfnt_r = math.log10((sumdoc-df)/sumdoc);
        sig1 = pt*(-idft_r);
        sig2 = pnt*(-idfnt_r);
        sig = sig1+sig2;
    return sig;

main();
```

Code for corpus signature extraction

```
import json
import math
import os
import string

def main():
    corpus_path = "H:\\Dropbox\\UNC\\MasterPaper\\Databridge\\corpus.json";
```

```
    # this corpus contains the document frequency of each term
    with open(corpus_path) as corpus_file:
        corpus = json.load(corpus_file);
    idf_list = [];
    for key in corpus.keys():
        idf = math.log(1+1000/float(corpus[key]));
        if idf<1.6:   #1.25 for 50 words #1.6 for 99 words
            idf_list.append(key);
    print len(idf_list);
    print idf_list;

    #print float(1000/1000)*math.log(float(1001)); # this is the case when the term is very unique 9.21
    #print float(1/1000)*math.log(float(2)); #this is the case when the term is most common 0.69

main();
```